UNIVERSIDADE DE BRASÍLIA
CURSO DE ESTATÍSTICA
TRABALHO DE CONCLUSÃO DE CURSO

TAÍSSA DE LIMA SANCHES

ANÁLISE DO RISCO DE INADIMPLÊNCIA DOS ESTUDANTES BENEFICIADOS PELO FUNDO DE FINANCIAMENTO ESTUDANTIL (FIES)

Brasília - DF
2018
TAISSA DE LIMA SANCHES

ANÁLISE DO RISCO DE INADIMPLÊNCIA DOS ESTUDANTES BENEFICIADOS PELO FUNDO DE FINANCIAMENTO ESTUDANTIL (FIES)

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Jhames Matos Sampaio.
Co-orientador: Prof. Andrea Felippe Cabello.

Brasília - DF

2018

TAÍSSA DE LIMA SANCHES

ANÁLISE DO RISCO DE INADIMPLÊNCIA DOS ESTUDANTES BENEFICIADOS PELO FUNDO DE FINANCIAMENTO ESTUDANTIL (FIES)

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Bacharel em Estatística.

Brasília, DF, ____ de _____ de _____.

Banca examinadora:

_____
Prof. Jhames Matos Sampaio (Orientador)

_____
Prof. Juliana Betini Fachini Gomes

_____
Prof. Lucas Moreira

Dedico este trabalho primeiramente a Deus, por ser essencial em minha vida, meu porto seguro e meu mestre. Ele esteve comigo em cada etapa até aqui, não me deixando fraquejar.
Também dedico à minha família que sempre me deu apoio para eu conseguir chegar até aqui.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pois sem ele esse trabalho não seria possível e a ajuda de todas as pessoas que contribuíram de alguma forma para a elaboração e conclusão desse trabalho.

Agradeço a toda à equipe de gestão de carteira no FNDE, que sempre foram muito prestativos para tirar as minhas dúvidas sobre o programa e fornecer informações essenciais. Agradeço especialmente ao meu supervisor de estágio Yves Dumaresq Sobral que não só confiou em mim, como incentivou e ajudou na elaboração, proporcionando durante esse trabalho um estágio de muita pesquisa e desenvolvimento sobre o tema.

Agradeço ao meu orientador Jhames Matos Sampaio que aceitou me orientar com um tema que eu sugeri a ele e teve muita paciência durante a elaboração desse trabalho.

Agradeço aos meus familiares que me deram apoio em todos os sentidos e sempre estiveram torcendo pelo fim dessa etapa em minha vida

Agradeço a todos os meus colegas de faculdade que com colaboração conseguimos superar todos os obstáculos para que esse trabalho fosse concretizado.

RESUMO

A inadimplência é algo cada vez mais presente nas organizações, apesar de serem feitas análises de risco antes da concessão de crédito. Assim como uma empresa, o governo também deve fazer estas análises, a fim de garantir uma boa gestão e melhor controle sobre o risco de crédito. O Fundo de Financiamento Estudantil (Fies) é um fundo de natureza contábil vinculado ao Ministério da Educação e destinado à concessão de financiamento a estudantes de cursos superiores não gratuitos. O Fies segue o formato de empréstimos bancários convencionais, em que o saldo devedor é distribuído por um número preestabelecido de parcelas, calculadas de maneira a saldá-lo ao fim do prazo-limite de amortização, sendo que o valor de cada parcela independe da situação financeira do mutuário na data de vencimento. O presente trabalho propõe uma classificação em faixas de risco de inadimplência para estes estudantes, visando à melhoria da gestão do fundo e um melhor conhecimento sobre os financiados, com o objetivo de prever a inadimplência dos estudantes beneficiados, que ainda não começaram a pagar. Para tal, utiliza-se de duas técnicas distintas: Regressão Logística e Random Forest, e para avaliação dos modelos e decisão sobre o melhor para esses dados, foi utilizado a matriz de confusão, a curva ROC, medida Kappa, a acurácia, a sensibilidade e a especificidade. Através das comparações, o Random Forest se mostrou um modelo mais eficiente para esses dados do que o modelo de Regressão Logística. Conclui-se que a forma pela qual o Banco Central faz o provisionamento do valor que se espera não ser recebido, se mostrou longe da realidade observada nos dois modelos. Para qualquer forma de modelo escolhido percebeu-se uma grande quantidade de estudantes dos quais se espera que não honrem com a dívida, o que poderá gerar um déficit orçamentário com graves consequências, como o fim do programa.

Palavras-chave: Análise de riscos de crédito. Inadimplência. Fundo de Financiamento Estudantil. Regressão logística. Random Forest.

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Bacen	Banco Central do Brasil
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
FGDUC	Fundo de Garantia de Operações de Crédito Educativo
Fies	Fundo de Financiamento Estudantil
FNDE	Fundo Nacional de Desenvolvimento da Educação
FN	Falso-Negativo
FP	Falso-Positivo
IES	Instituição de ensino Superior
Prouni	Programa Universidade Para Todos
ROC	Receiver Operating Characteristic.
SisFies	Sistema informatizado do Fies
TCU	Tribunal de Contas da União
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

A inadimplência - não pagar uma dívida - está cada vez mais presente nas organizações devido a fatores como má concessão de crédito; situação econômica do país; legislação vigente; falta de adaptação das organizações à nova realidade do mercado; baixo poder aquisitivo da população, dentre outros.

Ao conceder o crédito, as empresas devem analisar o risco que poderá ocorrer em virtude da incerteza do recebimento da dívida, visto que não se sabe ao certo se o cliente irá cumprir seus deveres (ANDRADE, 2010).

Seguindo este ponto de vista, o governo deve fazer esta análise como qualquer empresa para garantir uma boa gestão e para que alguns programas de empréstimo não sejam extintos pela falta de planejamento e controle.

Este trabalho destina-se a propor uma classificação em faixas de risco de inadimplência para os estudantes financiados através do Fundo de Financiamento Estudantil (Fies), mirando a melhoria da gestão do fundo e um melhor conhecimento sobre os financiados.

No passado, utilizaram-se avaliações e critérios subjetivos para conceder crédito, porém o risco de concedê-lo pode ser mais bem controlado com o uso de instrumentos estatísticos e de sistemas multivariados, possibilitando a mensuração do risco de forma mais objetiva e com uma abordagem empírica que enfatiza a predição (ALTMAN; SAUNDERS, 1997).

Saber do impacto que o crédito realizado por instituições públicas pode gerar na economia da sociedade, assim como a inadimplência nas finanças públicas, tendo em consideração que as peculiaridades do mercado de crédito educacional limitam o acesso à educação superior e justificam a adoção de políticas públicas de financiamento (ROCHA; EHRL; MONASTERIO, 2016) motiva a necessidade de analisar os estudantes que já adquiriram crédito estudantil com o governo.

Primeiramente, dar-se-á a contextualização do histórico do Fies; a partir desta etapa, justificar-se-á a necessidade da previsão no que diz respeito à inadimplência e suas implicações em relação aos estudantes beneficiados pelo Fies.

1.2 O FIES

O Fies é um fundo de natureza contábil, vinculado ao Ministério da Educação, destinado à concessão de financiamento a estudantes de cursos superiores não gratuitos e com avaliação positiva nos processos conduzidos pelo referido Ministério, de acordo com regulamentação própria (BRASIL, 2001).

Este, assim como a maioria dos programas de crédito educativo existentes no mundo, segue o formato de empréstimos bancários convencionais em um ponto central: o saldo devedor é distribuído por um número preestabelecido de parcelas, calculadas de maneira a saldá-lo ao fim do prazo-limite de amortização. O valor de cada parcela independe, pois, da situação financeira do mutuário na data de vencimento.

A diferença primordial entre linhas de crédito estudantis garantidas pelo governo e as linhas de crédito para as mais diversas finalidades oferecidas por instituições financeiras no mercado privado são os subsídios públicos (NASCIMENTO; LONGO, 2016). Estes subsídios garantem uma taxa de juros menor, além disso, que estudantes, os quais não receberiam empréstimos privados por terem um risco de inadimplência muito alto, consigam esse financiamento.

Tentando melhorar a taxa de adimplência e diminuir o número de fraudes que aconteceram em relação ao programa de financiamento, este foi reformulado em diversos pontos por várias vezes. Considera-se que a inadimplência é, potencialmente, prejudicial para a continuidade do programa, tendo em vista que são necessários infinitos aportes de recursos para que seja possível beneficiar outras pessoas.

1.3 JUSTIFICATIVA

O Fies tem passado por sucessivos ajustes em seu desenho após meia década de crescimento exponencial de seu orçamento e dos contratos de financiamento que o viabilizaram.

Tais ajustes buscam adequá-lo ao cenário de maior restrição fiscal no qual mergulhou o Brasil nos últimos dois anos e que não mais sustenta programas governamentais de financiamento que busquem, ao mesmo tempo, tão largo alcance e tão elevados subsídios, como foi o caso do Fies nos anos recentes (NASCIMENTO, LONGO; 2016).

Segundo Dynarski (2014), 40 milhões de pessoas nos Estados Unidos detêm débitos estudantis que já ultrapassam US$ 1 trilhão. Recentes crises de financiamento estudantil têm

aflorado e até mesmo chegaram a colocar governos em situações difíceis, como ocorridos no Chile e na Colômbia (SALMI, 2013), e os programas de crédito educativo em operação nesses países são menos subsidiados do que o do Brasil (NASCIMENTO; LONGO, 2016).

O Tribunal de Contas da União (TCU) já fez uma auditoria indicando a necessidade de avaliar a sustentabilidade do Fies, bem como a sua eficácia, as vulnerabilidades de seus processos de trabalho, sobretudo considerando-se a evolução expressiva no quantitativo de contratos formalizados nos últimos anos - a qual foi acompanhada da crescente demanda orçamentária para fazer frente às despesas do programa - além de avaliar-se o orçamento e a capacidade financeira deste programa de crédito estudantil visando à sua sustentabilidade, identificando eventuais riscos que possam impactar a continuidade dele (TCU, 2016.).

Considerando os problemas até aqui expostos, tem-se em vista a importância fundamental do crédito para o desenvolvimento da economia, através do provimento de recursos financeiros para que consumidores possam realizar seus projetos e adquirir bens com uma quantidade de dinheiro emprestado por uma instituição financeira, que deve ser reembolsado, com juros e em parcelas, proporcionando além de um crescimento econômico, a melhoria na qualidade de vida das pessoas (GERTLER; KARADI, 2015), segundo Chapman (2006), o investimento em crédito educacional apresenta riscos adicionais como:

a) Os estudantes matriculados não conhecem completamente a sua capacidade;
b) Eles não estão cientes da probabilidade de sucesso em sua área de estudo;
c) Existe incerteza quanto ao valor futuro do investimento, ou seja, o que parecia um bom investimento no início pode não o ser ao fim do curso;
d) Principalmente estudantes desfavorecidos socialmente podem não ter informação sobre os resultados possíveis da conclusão do nível superior.

Frente a esses motivos, é provável que os estudantes avessos ao risco sejam relutantes em buscar no mercado financeiro privado empréstimos/créditos educacionais e, ao mesmo tempo, os bancos privados considerem esses riscos e restrinjam o crédito para tal objetivo.

St. John et al. (2000) enfatiza que a análise custo-benefício do aluno tem componentes cognitivos (tangíveis) e afetivos (intangíveis). O componente cognitivo se concentra nos cálculos de custo e benefício, como é típico dos modelos de escolha racional. O componente afetivo inclui satisfação com a capacidade de pagar pela faculdade: “[...] incorpora o

estudante percepções sobre suas circunstâncias financeiras" (CABRERA et al., 1990 apud ST. JOHN et al., 2000, p. 37)[1].

As questões relacionadas ao número de dependentes, custo de vida do local onde reside e disposição para, realmente, honrar seus débitos são fatores que podem aumentar ou reduzir o peso dos encargos de reembolso para o indivíduo conjuntamente com os riscos de crédito associados (NASCIMENTO; LONGO, 2016). Pode-se concluir que classificar esse risco de inadimplência é de suma importância tanto para o governo quanto para o estudante, tendo em vista que incorporações de aspectos prudenciais podem tornar os programas de crédito estudantis mais eficazes.

Há embasamento teórico que mostra a importância de entender-se melhor alguns pontos desse programa do governo, tendo como base estudos que fazem parte da literatura referente ao Fies, tais como: Análise de impacto do Fies sobre o salário do trabalhador formal (ROCHA; EHRL; MONASTERIO, 2016); Qual foi o impacto do Fies nos salários? (ROCHA; MONASTERIO; EHRL, 2016); Qual o peso dos encargos de reembolso sobre a renda esperada dos beneficiários do Fies? (NASCIMENTO; LONGO, 2016).

[1] CABRERA, Alberto. F.; STAMPEN, Jacob; HANSEN, W. Lee. Exploring the effects of ability to pay on persistence in college. **Review of Higher Education**, Baltimore, v. 13, n. 3, p. 303–336, 1990.

3 OBJETIVOS

3.1 OBJETIVO GERAL

Analisar a inadimplência dos estudantes de todo o Brasil beneficiados pelo Fies, que ainda não começaram a pagar, com base na Regressão Logística e no Random Forest.

3.2 OBJETIVOS ESPECÍFICOS

a) Implementar um modelo de Regressão Logística para classificar o risco de adimplência e inadimplência;
b) Implementar um modelo Random Forest para classificar o risco de adimplência e inadimplência;
c) Comparar os dois modelos;
d) Aplicar os modelos no banco de dados dos estudantes que adquiriram crédito estudantil através do Fies e ainda não começaram a pagar.

4 TÉCNICAS ESTATÍSTICAS

As três principais técnicas para elaboração de sistemas de *Credit Scoring* são a análise discriminante, a Regressão Logística e, mais recentemente, modelos baseados em aprendizagem de máquina. Todas essas técnicas geram bons modelos e precisam manter-se atualizadas a fim de que se obtenham sempre resultados confiáveis.

Decidiu-se utilizar a técnica de Regressão Logística, como um exemplo de modelo clássico, por ser uma técnica há vários anos aceita pelos pesquisadores da área, além de mais robusta que a Análise Discriminante, não necessita satisfazer a suposição de normalidade dos dados e de que as matrizes de covariância sejam iguais para descobrir o valor ideal. Mesmo que esses pressupostos não sejam satisfeitos, a Regressão Logística ainda pode fornecer uma precisão relativamente alta de previsão (PRESS; WILSON, 1978).

Também se decidiu utilizar o Random Forest como um exemplo dos métodos de aprendizagem de máquina, por ele ser um método novo que tem apresentado bons resultados. Mostrando-se consonante com o trabalho "predição do bom e do mau pagador no programa minha casa, minha vida" do José Rômulo de Castro Vieira (2016), em que ele compara vários métodos para um mesmo banco de dados e o método que se mostra mais adequado é o Random Forest. Este presente trabalho também compara técnicas estatísticas de regressão e o o Random Forest se mostrando melhor.

Deve-se ressaltar que não se pode selecionar um algoritmo e reivindicar sua superioridade sobre algoritmos concorrentes sem ter em conta os dados, as características do problema, bem como a adequação do algoritmo para esses dados. No entanto, pode-se possivelmente reivindicar a superioridade de um algoritmo para um conjunto de dados ou problema específico (PIRAMUTHU, 2006).

Para avaliação desses modelos e averiguação de uma adequação suficiente para esses dados, utilizou-se a matriz de confusão, a curva ROC, medida Kappa, a acurácia, a sensibilidade e a especificidade nos dois casos.

4.1 REGRESSÃO LOGÍSTICA

A Regressão Logística é um método bastante tradicional e popular para a análise de *Credit Scoring*, seu uso deve estar de acordo com alguns critérios, por exemplo, afastar autocorrelação nos resíduos para evitar multicolinearidade nas variáveis independentes.

Este método permite estabelecer uma relação entre a variável resposta e as variáveis explicativas, sendo que a variável resposta é binária e as dependentes podem ser categóricas ou não. Os resultados da análise ficam contidos no intervalo de zero a um.

O Modelo Logístico apresenta vantagens como: facilidade para lidar com variáveis dependentes que são categóricas; fornece resultados em termos de probabilidade; facilidade de classificação de indivíduos em categorias; requer pequeno número de reposições e apresenta alto grau de confiabilidade.

Essa regressão estima uma relação linear entre variáveis e a probabilidade de pertencer a um grupo, neste trabalho, o indivíduo é classificado como adimplente ou inadimplente. A expressão do modelo logístico é apresentada da seguinte forma, sendo Y a variável resposta:

$$P(Y = 1) = \frac{1}{1 + e^{-Z}}, \tag{1}$$

Em que Z representa:

$$Z = b_0 + \sum_i^n b_i X_i, \tag{2}$$

$P(Y=1)$ e $P(Y=0)$ simbolizam a probabilidade de sucesso ou de fracasso respectivamente dependendo da variável X_i, sendo este a representação das variáveis explicativas.

Os coeficientes $b_0, b_1,..., b_n$ são estimados a partir do conjunto de dados, pelo método de máxima verossimilhança, em que encontra uma combinação de coeficientes que maximiza a probabilidade da amostra ter sido observada. Considerando certa combinação de coeficientes $b_0, b_1,..., b_n$ e variando os valores das variáveis regressoras. Observa-se que a Curva Logística, expressão gráfica da equação logística, tem um comportamento probabilístico no formato da letra S, o que é uma característica deste tipo de regressão. (HOSMER;LEMESHOW;1989).

Observa-se que, sendo: $Z(x) = b_0 + b_1 x_1 + ... + b_p x_p$

- $Z(x) \rightarrow +\infty$, então $P(Y = 1) \rightarrow 1$
- $Z(x) \rightarrow -\infty$, então $P(Y = 1) \rightarrow 0$

Para se interpretar os coeficientes, observa-se o impacto destes sobre a razão de chance-razão entre a chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo - depois se identifica o impacto do coeficiente da variável independente sobre a razão de chance; determina-se o efeito que os coeficientes exercem sobre a chance de um

evento ocorrer; ressalta-se que um coeficiente positivo aumenta a probabilidade e um negativo diminui a probabilidade.

4.2 RANDOM FOREST

Segundo Lantz (2015) o Random Forest baseia-se em conjuntos de árvores de decisões, combinando versatilidade e potência em uma abordagem de aprendizado de máquina única. O método consiste em um conjunto de árvores de decisões geradas dentro de um mesmo objeto. Cada objeto (conjunto de árvores) passa por um mecanismo de votação (*bagging*), que elege a classificação mais votada. Um exemplo de árvore de decisão interna ao classificador é exibido na Figura 1

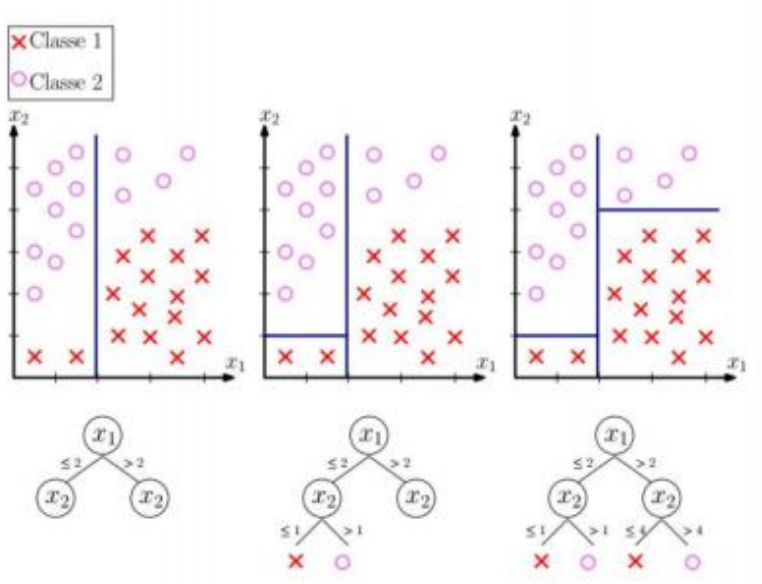


Figura 1: Fonte: (LOPEZ;2014)

Pode-se perceber pela Figura 1 que o classificador Random Forest separa as superfícies de decisão por meio da criação de uma sequencia de hiperplanos paralelos aos eixos.

Em geral um classificador particiona o espaço de características em volumes designados regiões de decisão. Todos os vetores de características no interior de uma região são atribuídos à mesma categoria. A região de decisão para uma classe pode ser simplesmente conexa, ou pode consistir em duas ou mais sub-regiões não adjacentes. As regiões de decisão encontram-se separadas por superfícies designadas superfícies de decisão ou superfícies de separação e estas superfícies representam postos onde existem “empates” entre duas ou mais categorias. (ASCENSO; FRED; 2001)

O Random Forest utiliza apenas uma pequena parte aleatória do conjunto completo de observações, podendo lidar com grandes conjuntos de dados; além de ser um classificador que consiste em uma coleção de árvores classificadoras estruturadas h(x; **θ**k); k = 1...; **θ** é parâmetro; os **θ**k são vetores aleatórios independentes, identicamente distribuídos e cada árvore lança um único voto para a classe mais popular a partir dos dados de entrada *x* (BREIMAN, 2001).

4.3 MATRIZ DE CONFUSÃO

A matriz de confusão é uma maneira de constatar se o modelo está prevendo corretamente os bons e maus clientes, a forma de interpretar os valores dessa técnica está definida na Tabela 1.

Tabela 1 – Matriz de confusão

	Real		
Previsto	**Mau**	**Bom**	**Total**
Bom	FP	VP	VP+FP
Mau	VN	FN	FN+VN
Total	FP+VN	VP+FN	VP+FP+FN+VN

Fonte: os autores (2018).

Em que os elementos dessa matriz são definidos como:

- VP: representa o número de maus pagadores, classificados corretamente como maus, ou seja, Verdadeiro Positivo;
- FP: representa o número de bons pagadores, classificados incorretamente como maus, ou seja, Falso-Positivo;
- FN: representa o número de maus pagadores, classificados incorretamente como bons, ou seja, Falso-Negativo;
- VN: representa o número de bons pagadores classificados corretamente como bons, ou seja, Verdadeiro Negativo.

4.4 MEDIDAS PREDITIVAS

As medidas preditivas são definidas com base nas quantidades da matriz de confusão. Essas medidas são a acurácia, sensibilidade e especificidade, descritas por:

$$Acurácia(AC) = \frac{\text{VP+VN}}{VP+VN+FP+FN}, \tag{3}$$

$$Sensibilidade(S) = 1 - Erro\ Tipo\ I = \frac{\text{VP}}{VP+FN}, \tag{4}$$

$$Especificidade(E) = 1 - Erro\ Tipo\ II = \frac{\text{VN}}{VN+FP}, \tag{5}$$

Em um teste de hipótese o Erro Tipo I consiste em rejeitar a hipótese nula quando esta é verdadeira. O Erro Tipo II é o erro que ocorre quando o teste de hipótese não rejeita a hipótese nula quando esta é falsa. A Acurácia representa a taxa de acerto do modelo. A Sensibilidade indica a capacidade do modelo de identificar os maus pagadores enquanto a Especificidade representa a capacidade de detectar aqueles que são bons pagadores.

4.5 KAPPA

Esse coeficiente serve para comparar dois ou mais "juízes" para um mesmo banco de dados. Ele apresenta valores negativos para quando o modelo é pior do que um modelo aleatório e mostra valor igual a 1 para uma situação ideal. Landis e Koch (1977) sugerem a seguinte interpretação para o Kappa (Tabela 2):

Tabela 2 – Interpretação de Kappa

Valores de Kappa	Interpretação
<0	Insignificante
0-0.19	Pobre
0.20-0.39	Fraca
0.40-0.59	Razoável
0.60-0.79	Forte
0.80-1.00	Quase perfeita

Fonte: Landis e Koch (1977).

O coeficiente de concordância de Kappa, sugerido por Cohen em 1960 é calculado utilizando a equação (6):

$$\widehat{K} = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \quad (6)$$

Considerando p_0 a taxa de aceitação relativa, p_e a taxa hipotética de aceitação.

4.6 RECEIVER OPERATING CHARACTERISTIC (ROC)

ROC é um método visual que pode ser construído a partir de duas amostras de escores, uma para casos atípicos, como devedores inadimplentes, e outra para casos regulares. A curva ROC é um teste que busca mostrar a relação, normalmente antagônica, entre a sensibilidade e a especificidade. A área da curva está relacionada com a distribuição de frequência de eventos de inadimplência e não inadimplência e permitem quantificar a exatidão de um teste, pois, quanto maior a área sob a curva ROC, maior é a precisão (ENGELMANN; HAYDEN; TASCHE, 2003). Ou seja, a curva ROC vai ser útil para comparar os modelos, sendo mais preciso o que apresentar uma maior área abaixo da curva.

A área varia de 0,5 para modelos aleatórios, sem poder discriminativo, até 1 para uma situação ideal. Para calcular a área da curva ROC é necessário conhecer o valor do score.

$$HR(C) = \frac{H(C)}{N_D}, \quad (8)$$

Neste caso H(C) é o número de inadimplentes com score inferior a C e N_D é o total de inadimplentes. A taxa de falsos alarmes FAR(C) é:

$$FAR(C) = \frac{F(C)}{N_{ND}}, \quad (9)$$

sendo que F(C) corresponde ao número de não inadimplentes com pontuações menores que o score e N_{ND} corresponde ao número total de não inadimplentes.

Pelas equações (8) e (9) é possível chegar ao valor da área da curva ROC com a seguinte equação:

$$\text{AUROC} = \int_0^1 HR(FAR)d(FAR) \quad (10)$$

4.7 INSTRUMENTOS ESTATÍSTICOS –R

O software R já vem com distribuições compiladas para Windows, Mac OS ou Linux além de ser gratuito oferece pacotes, também sem custo, que podem ser instalados dentro da própria plataforma para facilitar o uso de algumas técnicas.

Para empregar o Random Forest foram utilizados os pacotes “RandomForest” e “H2O”. O pacote “ROCR” foi utilizado para mostrar as curvas ROC e o cálculo da AUROC.

A biblioteca “H2O” foi utilizada para implementar o processamento em paralelo (multi-core), visto que o número de mais de 2 milhões de observações dificulta a implementação do Random Forest que sem esse pacote se torna lento e ineficiente.

5 METODOLOGIA

Este trabalho contém dois modelos de *Credit Scoring*: um tradicional (Regressão Logística) que é uma técnica amplamente difundida e utilizada pelos estudiosos da área há vários anos, e um de aprendizagem de máquina. Nos dois casos, a base de dados com as informações dos estudantes egressos das Instituições de Ensino Superior (IES), financiados pelo Fies, será dividida em "base de treino" e "base teste", sendo que a base de treino será utilizada para que se obtenham os modelos e a base de teste para validá-los.

Após a validação, estes modelos serão utilizados nos dados referentes aos estudantes que não iniciaram o pagamento da dívida.

Para esse estudo a variável "*default*" será calculada com base em parcelas vencidas a mais de 365 dias, classificadas como saldo perdido para o Fies.

5.1 CREDIT SCORING

Credit Scoring pode ser definido como o processo de atribuição de pontos às variáveis de decisão mediante técnicas estatísticas. Trata-se de processo que define a probabilidade de um cliente com certas características pertencer ou não a um grupo possuidor de outras determinadas características consideradas desejáveis, sendo o caso que se aprova um limite de crédito. Essa classificação estabelece uma regra de discriminação de um determinado cliente solicitante de crédito (VICENTE, 2001).

5.2 BANCO CENTRAL DO BRASIL (Bacen)

A Resolução nº 2682/1999 do Banco Central (Bacen) estabeleceu que as instituições financeiras devem classificar sua exposição de crédito em nove níveis de risco de acordo com o sistema de classificação da Tabela 3

Tabela 3 – Classificação da inadimplência de acordo com o Banco Central do Brasil (1999)

	AA	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Provisão (%)	0	0,5	1	3	10	30	50	70	100
Níveis de atraso (dias)	-	-	15-30	31-60	61-90	91-120	121-150	151-180	>180

Fonte: Bacen (1999).

As provisões encontradas nos modelos serão comparadas com as que seriam indicadas pelo Bacen. Assim, tendo em vista que as provisões encontradas utilizando a Regressão Logística e o Random Forest são calculadas especificamente para os dados do Fies e as mesmas que seriam indicadas pelo Bacen (1999) são generalizadas para todas as situações, e baseadas apenas nos dias de atraso, espera-se que as provisões encontradas nos modelos estatísticos deste trabalho sejam mais fidedignas com a realidade.

5.3 JUSTIFICATIVA

Os sistemas de pontuação de crédito são encontrados em praticamente todos os tipos de análises de crédito, desde crédito ao consumidor até empréstimos comerciais. A ideia é essencialmente a mesma: a pré-identificação de certos fatores-chave que determinam a probabilidade de inadimplência, e sua combinação ou ponderação para produzir uma pontuação quantitativa (SAUNDERS, 2000).

Parkinson e Ochs (1998, p. 26-27) elaboraram um resumo com as principais vantagens do uso de sistemas de *Credit Scoring*:

a) **Revisões de crédito consistente**: os dados históricos de outros devedores são um bom indicador de consistência para revisão de crédito;
b) **Informações organizadas**: a sistematização e organização das informações contribuem para a melhoria do processo;
c) **Eficiência no trato de dados fornecidos por terceiros**: o processo de *Credit Scoring* torna as informações de banco de dados fornecido por terceiros, anteriormente classificadas como dados acessórios, parte integrante do sistema;
d) **Diminuição da metodologia subjetiva**: o uso de um sistema quantitativo parametrizado que minimiza o subjetivismo;
e) **Compreensão do processo**: o modelo construído sintetiza o processo de concessão de crédito da empresa, fornecendo maiores subsídios para entendê-lo;
f) **Maior eficiência do processo**: a análise de crédito é centrada em um número menor de fatores, reduzindo o tempo do processo e melhorando a eficiência.

5.5 DADOS

O Banco de dados formou-se pelos bancos de dados utilizadas para pagamento do Banco do Brasil e da Caixa Econômica Federal e dos sete bancos de dados do Sistema Informatizado do Fundo de Financiamento Estudantil (SisFies) com informações sobre os abatimentos, alunos, cursos, fiadores, financiamentos, instituições de ensino e mantenedoras. A equipe de gestão de carteira do Fies no Fundo Nacional de Desenvolvimento da Educação (FNDE) consolidou todas essas informações, por meio do *software* SQL, e utilizou-se esse banco de dados.

A pesquisa realizada utiliza a tabela consolidada do mês de setembro de 2017, cada linha da tabela corresponde a um aluno e cada coluna uma correspondente informação, totalizando-se 2.551.852 linhas e 45 colunas. Para realizar-se a classificação foram usados apenas os egressos do Fies que apresentavam o contrato nas fases de amortização ou liquidado de acordo com o SisFies, consolidando um banco de dados com 594.772 linhas empregadas para construção e validação do modelo.

Outras variáveis foram criadas com base em informações existentes como, por exemplo, a idade do estudante na data da assinatura do contrato e a área de conhecimento, que foi classificada associando-se informações do curso de acordo o manual do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Com base nos artigos "Qual foi o impacto do Fies nos salários?" de Rocha, Monasterio e Ehrl (2016) e no artigo "*The effect of loans on the persistence and attainment of community college students*" de Dowd e Coury (2006) obteve-se embasamento para se considerar tanto variáveis ambientais, como variáveis sociais em um primeiro momento, ou seja, todas as informações disponíveis sobre os alunos, cabíveis para o modelo, foram consideradas para depois efetuar-se uma seleção de variáveis e saber quais realmente influenciam na inadimplência e quais têm correlação entre si.

Para o modelo de Regressão Logística todas as variáveis foram categorizadas de acordo com as Tabelas 22- 33 (univariadas) e Tabelas 45- 56 (bivariadas). No Random Forest as variáveis foram aplicadas direto ao modelo sem categorização.

5.7 PROCEDIMENTOS

Antes de utilizar alguma técnica estatística padronizou-se o banco de dados, por exemplo, no momento de preencher se o aluno fez ensino médio em escola pública ou privada alguns responderam no caso de escola pública “S” correspondendo a “sim” e outros responderam “P” equivalente a “pública” em contraponto aos que estudaram em escola privada no ensino médio que estavam classificados como “N”. Nesse primeiro momento realizou-se um padrão de classificação, sendo “S” para os que estudaram em escola pública e “N” para os que estudaram em escolas privadas.

Por tratar-se de 26 unidades da federação, além do Distrito Federal achou-se por bem separa-las por regiões geográficas correspondentes a localização dos estudantes e IES. Além disso foi estabelecida uma variável correspondente às oito áreas de conhecimento conforme classificação do CNPq.

Ao se utilizar variáveis “*dummy*” obtém-se *n*-1 variáveis, neste caso, seriam 26 variáveis correspondentes às unidades da federação tendo em vista a existência de IES e estudantes em todo território nacional, o que tornaria sua utilização inviável e por isso trabalhou-se com as regiões do Brasil. Também se obtiveram 38.743 códigos de cursos inscritos no Fies, essa dificuldade relativa às variáveis *dummy*, é encontrada principalmente para a Regressão Logística. No Random Forest é mais simples lidar com essa quantidade de opções para a mesma variável, pois ele cria árvores de decisões internas ao invés de variáveis *dummy*.

Percebe-se que as variáveis correspondentes ao valor da renda comprometida dos fiadores, valor da renda comprovada dos fiadores e valor da renda informada dos fiadores apresentam muitos valores faltantes. Esses “*no answer”* (Na’s) se sucederam do fato de que muitos estudantes não têm nenhuma pessoa física como fiador, por terem uma renda compatibilizada para a única garantia ser o Fundo de Garantia de Operações de Crédito Educativo (FGDUC).

Os estudantes do Mato Grosso atualmente têm uma liminar na justiça para não precisarem apresentar fiadores, os estudantes de Alagoas e do Rio Grande do Norte que apesar de atualmente não terem mais essa prerrogativa, para a não necessidade de fiador, podem não ter feito nenhum aditamento depois do fim dessa liminar e começaram a amortizar ou liquidaram o contrato sem fiador.

A idade dos estudantes, na data de assinatura do contrato, foi calculada de acordo com a data de nascimento, construindo uma nova variável chamada “idade”.

Algumas variáveis, que mostravam a mesma informação que outra ou que não mostravam relevantes para esse estudo foram excluídos, como por exemplo, no caso em que apareciam as variáveis “banco” e “código do banco” apenas a variável “banco” foi mantida, visto que as duas variáveis mostravam exatamente a mesma informação. As variáveis excluídas foram “CodigoBanco”, “CpfAluno”, “Cep”, “CidadeAluno”, “FaseBanco”, “FaseSISFIES”, “Risco”, “NuSemestreReferencia”, “SafraEntradaSISFIES”, “SafraSaidaSISFI”, “SafraEntradaAF”, “CodCurso”, “CodTurno”, “MunicipioIes”, “CnpjIES”, “CnpjMant”, “Contagem”, “DtNascimento”, “DtAssinaturaContratoAF”, “ValorDividaAF”, “ValorLiberadoAF” e “Resposta”.

Como a quantidade de alunos é muito grande não se mostrou um problema retirar alguns dados que apresentavam *outliers* e provavelmente existiam por causa de algum erro. Exclui-se dessa análise alunos que apresentavam à renda familiar mensal bruta e/ou a renda pessoal mensal bruta maior que R$ 20.000,00, além dos que apresentavam idade menor do que 15 anos e/ou maior do que 80 anos.

Ao estudante inadimplente, que atrasa o contrato mais de 365 dias, considerou-se como pertencente à categoria 1 na variável resposta e ao estudante adimplente considerar-se-á a categoria 0.

A modelagem será feita em quatro partes que consistem na análise descritiva dos dados, construção, validação e aplicação dos modelos.

Todos os procedimentos realizar-se-ão na plataforma RStudio versão 1.1.423

6 ANÁLISE DE DADOS

A seguir apresentam-se as Tabelas 4-56 que foram usadas análise descritiva. Todos esses dados são referentes aos alunos que estavam nas fases de amortização ou com o contrato liquidado em setembro de 2017, pois esses dados que foram utilizados para se obter o modelo.

6.1 MEDIDAS RESUMO DAS VARIÁVEIS QUANTITATIVAS

As Tabelas 4-10 a seguir apresentam o valor mínimo, o 1° Quartil, a mediana, a média, o 3° Quartil, o valor máximo e a quantidade de valores não informados, quando isso ocorrer, para as variáveis quantitativas: renda familiar mensal bruta, renda pessoal mensal bruta, valor financiado de acordo com o SisFies, soma do valor da renda comprovada dos fiadores, taxa de juros, idade e o percentual financiado.

Tabela 1 – Medida descritiva da variável renda familiar mensal bruta dos estudantes egressos do Fies de todo o Brasil até setembro de 2017

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
0	950	1.591	2.140	2.643	20.000

Fonte: os autores (2018).

A Tabela 4 mostra que a média em relação a variável renda familiar bruta é maior que a mediana e o valor máximo encontrado foi equivalente ao maior valor de renda familiar bruta valor aceito para se ter um financiamento estudantil pelo Fies.

Tabela 2 – Medida descritiva da variável renda pessoal mensal bruta dos estudantes egressos do Fies de todo o Brasil até setembro de 2017

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
0	0	622	663,4	976	20.000

Fonte: os autores (2018).

A Tabela 5 mostra que a média em relação a variável renda mensal bruta próximo da mediana e o valor máximo encontrado foi equivalente ao maior de renda pessoal bruta do valor aceito para se ter um financiamento estudantil pelo Fies.

Tabela 3 – Medida descritiva da variável valor financiado SISFIES dos estudantes egressos do Fies de todo o Brasil até setembro de 2017

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
0	7.662	14.339	21.096	25.354	17348803

Fonte: os autores (2018).

A Tabela 6 mostra que a variável valor financiado SISFIES está bem distribuído existindo estudantes com valores muito altos, o valor é igual à zero (provavelmente nunca cursaram algum curso com o financiamento) e o 1° quartil, média, mediana e o 3° quartil apresentam valores crescentes.

Tabela 4 – Medida descritiva da variável soma do valor da renda comprovada dos estudantes fiadores dos egressos do Fies de todo o Brasil até setembro de 2017

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo	Não informado
0	1.708	2.687	4.228	4.662	1.872.64	312.869

Fonte: os autores (2018).

A Tabela 7 apresenta um valor máximo referente à soma do valor da renda comprovada dos fiadores elevado em relação ao 3° quartil, sendo este com um valor próximo ao valor referente ao valor da média.

Tabela 5 – Medida descritiva da variável taxa de juros dos estudantes egressos do Fies de todo o Brasil até setembro de 2017

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
3,4	3,4	3,4	3,4	3,4	6,5

Fonte: os autores (2018).

A Tabela 8 mostra que a grande maioria dos estudantes egressos do Fies de todo o Brasil até setembro de 2017 pagaram uma taxa de juros de 3,4% e uma pequena parte financiou com 6,5% de juros o curso.

Tabela 6 – Medida descritiva da variável idade dos estudantes egressos do Fies de todo o Brasil até setembro de 2017

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
16	20	23	25.28	29	79

Fonte: os autores (2018).

A Tabela 9 mostra que apesar de existirem estudantes egressos do Fies de todo o Brasil até setembro de 2017 com idade entre 16 e 79 anos a maioria assinou o contrato de financiamento quando tinham entre 20 e 29 anos de idade.

Tabela 7 – Medida descritiva da variável percentual financiado dos estudantes egressos do Fies de todo o Brasil até setembro de 2017

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
5%	75%	100%	89,08%	100%	100%

Fonte: os autores (2018).

A Tabela 10 mostra que mais da metade dos egressos do Fies de todo o Brasil até setembro de 2017 obtiverem financiamento igual a 100% do valor do curso, que o menor percentual financiado for de 5% e que o percentual médio financiado foi de 89,08%.

6.2 TABELAS UNIVARIÁDAS SEM CATEGORIZAÇÃO

AsTabelas 11-21 são univariadas das variáveis que correspondem: ao tipo de garantia, ao banco, ao sexo do aluno, ao tipo de escola onde cursou o ensino médio, se o aluno é portador de necessidades especiais, se o aluno é beneficiário do Programa Universidade Para Todos (Prouni) concomitantemente, situação da inscrição do Fies, região da instituição de ensino contratada, estado civil do aluno, turno do curso contratado e a área de conhecimento do curso.

Tabela 8 – Percentual de estudantes em cada tipo de garantia dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Tipo de garantia	Percentual (%)
Fgduc	49,20
Fiança convencional	45,73
Fiança solidária	1,67
Fiança convencional+ Fgduc	3,23
Fiança solidária + Fgduc	0,05
Sem fiança	0,078

Fonte: os autores (2018).

A Tabela 11 mostra que a maioria dos estudantes (49,20%) optou pela garantia do Fgduc, seguido pela fiança convencional que abrange 45,75% dos estudantes egressos e essas duas modalidades de garantia totalizam 94,94% dos estudantes.

O menor percentual dos estudantes utilizou a fiança solidária sendo equivalente a 0,05% do total dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Tabela 9 - Percentual de estudantes em cada banco contratado dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Banco	Percentual (%)
Banco do Brasil	38,65
Caixa Econômica Federal	61,34

Fonte: os autores (2018).

A Tabela 12 mostra que a Maioria dos estudantes paga o financiamento estudantil pela Caixa Econômica Federal (61,34%) e o restante dos estudantes pagam pela Banco do Brasil(38,65%).

Esses são os dois únicos banco disponíveis para firmar o contrato de financiamento e provavelmente mais alunos são utilizam a Caixa Econômica Federal pois no início do programa apenas este banco poderia ser utilizado para ter-se um contrato com o Fies.

Tabela 10 – Percentual de estudantes de cada sexo dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Sexo	Percentual (%)
Feminino	59,44
Masculino	40,55

Fonte: os autores (2018).

A Tabela 13 mostra que 59,44% dos estudantes egressos do Fies de todo o Brasil até setembro de 2017 eram mulheres e 40,55% eram homens.

Tabela 11 – Percentual de estudantes em cada tipo de escola onde cursou o ensino médio dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Escola	Percentual (%)
Pública	82,19
Privada	17,80

Fonte: os autores (2018).

A Tabela 14 mostra que 82,19% dos estuda beneficiados pelo programa cursaram o ensino médio em escola pública, o que realmente está de acordo com as intenções do Fies, e 17,80% dos beneficiados até Setembro de 2017 cursaram ensino médio em escolas privadas.

Tabela 12 – Percentual de estudantes portador de necessidade especial dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

PNE	Percentual (%)
Sim	0,87
Não	99,12

Fonte: os autores (2018).

A Tabela 15 mostra que a quantidade de estudantes portadores de necessidades especiais que obtiveram financiamento até Setembro de 2017 foi muito pequena para utilizar-se essa variável para inferir se ela influencia na inadimplência.

Tabela 13 – Percentual de estudantes que também fazem uso do Prouni dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Prouni	Percentual (%)
Não	95,74
Sim	4.25

Fonte: os autores (2018).

A Tabela 16 mostra que 95,74% dos estuda beneficiados pelo programa não obtiveram paralelamente bolsa de estudos fornecida pelo Prouni e 4,25% dos beneficiados até Setembro de 2017 obtiveram uma bolsa Parcial do Prouni.

Tabela 14 – Percentual de estudantes em cada situação da inscrição dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Situação	Percentual (%)
Contratado	84,93
Contrato encerrado	15,06

Fonte: os autores (2018).

A Tabela 17 mostra que 15,06% dos estudantes encerraram o contrato para entrar na faze de amortização.

Tabela 15 – Percentual de estudantes em cada região da instituição de ensino dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Instituição de Ensino	Percentual (%)
Centro-Oeste	11,36
Nordeste	18,75
Norte	4,67
Sudeste	50,10
Sul	14,91
Não informado	0,18

Fonte: os autores (2018).

A Tabela 18 mostra que 50,10% dos contratos em amortização e liquidados até Setembro de 22017 foram para o Sudeste, 18,75% foram para o Nordeste, 14,91% foram para o Sul, 11,36% foram para p Centro-Oeste, 4,67% foram para o Norte e 0,18% não informaram a região em que se encontrava a instituição de ensino.

Tabela 16 – Percentual de estudantes em cada estado civil dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Estado civil	Percentual (%)
Casado	15,32
Divorciado	2,50
Separado	1,16
Solteiro	77,77
União Estável	3,04
Viúvo	0,18

Fonte: os autores (2018).

A Tabela 19 mostra que 77,77% dos contratos em amortização e liquidado até Setembro de 2017 foram para estudantes solteiros, 15,32% foram para estudantes casados, 3,04% foram para estudantes em União Estável, 2,5% foram para estudantes Divorciados, 1,16% foram para estudantes Separados e 0,18% fora para estudantes Viúvos.

Tabela 17 – Percentual de estudantes em cada turno dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Turno	Percentual (%)
Integral	5,5
Matutino	18,41
Não aplica	0,00067
Noturno	73,50
Vespertino	2,57

Fonte: os autores (2018).

A Tabela 20 mostra que o período do curso dos estudantes egressos do Fies até setembro de 2017 foi: 73,50% para o período noturno, 18,41 no período Matutino, 5,5 no período integral, 2,57% no período vespertino e 0,00067 não aplica.

Tabela 18 – Percentual de estudantes em cada área de conhecimento dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Área de conhecimento	Percentual (%)
Educação	10,28
Humanidade e Artes	4,11
Ciências Sociais, Negócios e Direito.	40,77
Ciências, Matemática e Computação.	8,28
Engenharia, Produção e Construção.	12,28
Agricultura e Veterinária.	1,72
Saúde e bem estar social.	18,47
Serviços	3,34

Fonte: os autores (2018).

A Tabela 21 mostra que a porcentagem de financiamentos de acordo com a área de conhecimento do curso dos estudantes egressos do Fies até setembro de 2017 foi: 40,77% para Ciências Sociais, Negócios e Direito; 18,47% para Saúde e bem estar social; 12,28% para Engenharia, Produção e Construção; 10,28% para Educação; 4,11 para Humanidade e Artes; 3,34 para Serviços e; 1,72% para Agricultura e Veterinária.

6.3 TABELAS BIVARIDAS SEM CATEGORIZAÇÃO

As Tabelas 22-26 bivariadas mostram a quantidade de estudantes adimplentes e inadimplentes para cada opção possível dentro das variáveis. Utilizaram-se essas tabelas para categorizar os estudantes por isso construiu-se essa categoria de tabelas apenas para com mais do que duas opções em cada variável.

Tabela 22 – Adimplência e inadimplência em relação à região do Brasil que se encontra a instituição de ensino dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Região	Adimplente (%)	Inadimplente (%)
Centro Oeste	64,93	30,06
Nordeste	74,19	25,80
Norte	68,47	31,52
Sudeste	73,93	26,06
Sul	89,85	10,14
Não informado	49,86	50,13

Fonte: os autores (2018).

Observando-se o percentual de inadimplência da Tabela 22 percebe-se que se pode agrupar em três categorias sendo a categoria A com as regiões: não informado, Nordeste e Sudeste; a categoria B com Centro Oeste e Norte; e a categoria C com a região Sul.

Tabela 23 – Adimplência e inadimplência em relação ao estado civil do estudante dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Estado civil	Adimplente (%)	Inadimplente (%)
Casado	72,77	27,22
Divorciado	70,69	29,30
Separado	69,46	30,53
Solteiro	75,90	24,09
União Estável	70,32	29,67
Viúvo	70,27	29,72

Fonte: os autores (2018).

Observando-se o percentual de inadimplência da Tabela 23 percebe-se que se pode agrupar em três categorias sendo a categoria A com os estados civis: Divorciado, Separado, União estável e viúvo; a categoria B com o estado cível Solteiro; e a categoria C com o etado cível Casado.

Tabela 24 – Adimplência e inadimplência em relação ao turno dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Turno	Adimplente (%)	Inadimplente (%)
Integral	91,74	8,25
Matutino	75,17	24,82
Noturno	73,54	26,45
Vespertino	80,75	19,24

Fonte: os autores (2018).

Observando-se o percentual de inadimplência da Tabela 24 percebe-se que se pode agrupar em duas categorias sendo a categoria A com os turnos: Integral e não Aplica; e a categoria B com os turnos matutino, noturno e vespertino.

Tabela 25 – Adimplência e inadimplência em relação à garantia do financiamento dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Garantia	Adimplente (%)	Inadimplente (%)
Fgduc	60,49	39,50
Fiança convencional	90,06	9,93
Fiança solidaria	54,52	45,47
Fiança convencional+ Fgduc	94,57	5,42
Fiança solidaria + Fgduc	62,10	37,89
Sem Fiança	71,36	28,63
Não informado	87,85	12,14

Fonte: os autores (2018).

Observando-se o percentual de inadimplência da Tabela 25 percebe-se que se pode agrupar esta variável em três categorias sendo a categoria A com as garantias: Fiança convencional, não informado, Fiança convencional com Fgduc; a categoria B com as garantias Fgduc e Fiança solidária com Fgduc; e a categoria C com as garantias Fiança Solidária e Sem Fiança.

Tabela 26 – Adimplência e inadimplência em relação à área de conhecimento dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Área de conhecimento	Adimplente (%)	Inadimplente (%)
Educação	66,33	33,66
Humanidade e Artes	67,98	32,01
Ciências Sociais, Negócios e Direito	72,74	27,25
Ciências, Matemática e Computação	76,31	23,68
Engenharia, Produção e Construção	85,25	14,74
Agricultura e Veterinária	85,50	14,49
Saúde e bem estar social	79,73	20,26
Serviços	64,15	35,84

Fonte: os autores (2018).

Observando-se o percentual de inadimplência da Tabela 26 percebe-se que se pode agrupar esta variável em quatro categorias sendo a categoria A com as áreas de conhecimento: Educação, Humanidade e Artes e Serviços; a categoria B com as áreas de conhecimento Ciências Sociais, Negócios e Direito, Ciências, Matemática e Computação e Saúde e bem estar social; a categoria C com as áreas de conhecimento Engenharia, Produção e Construção e Agricultura e Veterinária e a categoria D os estudantes que não informaram a área de conhecimento.

6.5 TABELAS BIVARIDAS CATEGORIZADAS

A seguir estão as Tabelas 27-33 que são a categorização das variáveis contínuas e por esse motivo a categorização não se fez da mesma maneira das Tabelas 22-26 que apresentavam variáveis discretas.

Para as variáveis correspondentes as das Tabelas 22-26 a categorização foi feita de acordo com as bivariadas sem categorização juntando percentuais próximos de inadimplência e para a categorização das variáveis das Tabelas 27-33 foram feitas observando os percentuais de inadimplência de acordo com vários cortes nas variáveis contínuas.

As Tabelas 27-33 mostram a faixa de categorização para as variáveis correspondente a quantidade de semestres contratados, renda familiar mensal bruta, renda pessoal mensal bruta, valor financiado, soma da renda dos fiadores, idade e percentual financiado.

Percebe-se por essa categorização que a diferença de inadimplência de uma categoria para a outra procura ser a maior possível.

Tabela 27 – Adimplência e inadimplência em relação à quantidade de semestres categorizados dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Quantidade de semestres contratados categorizados	Categorias	Adimplente (%)	Inadimplente (%)
1 a 8	A	76,92	23,07
Mais do que 9 semestres	B	72,75	27,24

Fonte: os autores (2018).

Tabela 28– Adimplência e inadimplência em relação à renda familiar mensal bruta categorizada dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Renda familiar mensal bruta categorizada	Categorias	Adimplente (%)	Inadimplente (%)
Renda familiar mensal bruta ≤ R$ 1100,00	A	63,16	36,83
R$ 1100,00 < Renda familiar mensal bruta ≤ R$ 2500,00	B	75,63	24,36
Renda familiar mensal bruta > R$ 2500,00	C	88,43	11,56

Fonte: os autores (2018).

Tabela 29 – Adimplência e inadimplência em relação à renda pessoal mensal bruta categorizada dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Renda pessoal mensal bruta categorizada	Categorias	Adimplente (%)	Inadimplente (%)
Renda pessoal mensal bruta ≤ R$ 450,00	A	78,76	21,23
R$ 450,00 < Renda pessoal mensal bruta ≤ R$ 815,00	B	68,21	31,78
Renda pessoal mensal bruta > R$ 815,00	C	75,50	24,49

Fonte: os autores (2018).

Tabela 30 – Adimplência e inadimplência em relação ao valor financiado categorizado dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Valor financiado categorizado	Categorias	Adimplente (%)	Inadimplente (%)
Valor financiado ≤ R$ 7671,00	A	63,97	36,02
R$ 7671,00 < Valor financiado ≤ R$ 25395,00	B	74,91	25,08
Valor financiado > R$ 25395,00	C	86,39	13,60

Fonte: os autores (2018).

Tabela 31 – Adimplência e inadimplência em relação à soma da renda comprovada dos fiadores categorizado dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Valor financiado categorizado	Categorias	Adimplente (%)	Inadimplente (%)
Soma da renda dos fiadores ≤ R$ 1710,00	A	87,37	12,62
R$ 1710,00 < Soma da renda dos fiadores ≤ R$ 2692,00	B	90,27	9,72
Soma da renda dos fiadores > R$ 2692,00	C	94,26	5,73
Sem fiança convencional	D	60,21	39,78

Fonte: os autores (2018).

Tabela 32 – Adimplência e inadimplência em relação à idade categorizada dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Idade categorizada	Categorias	Adimplente (%)	Inadimplente (%)
Idade ≤ 20 anos	A	80,71	19,28
Idade > 20 anos	B	72,61	27,38

Fonte: os autores (2018).

Tabela 33 – Adimplência e inadimplência em relação ao percentual financiado categorizado dos estudantes egressos do Fies de todo o Brasil até setembro de 2017.

Percentual financiado categorizado	Categorias	Adimplente (%)	Inadimplente (%)
Financiamento = 100%	A	70,45	29,54
Financiamento < 100%	B	85,58	14,41

Fonte: os autores (2018).

7 CONSTRUÇÃO DOS MODELOS

7.1 CONSTRUÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA

As variáveis foram divididas em dois bancos de dados, um teste e um treino. O teste equivale-se de 30% dos dados com 178.432 observações e o treino com 70% dos dados com 416.340 observações.

Todas as variáveis, exceto curso e estudante portador de necessidade especial, foram categorizadas antes de serem utilizadas no modelo de Regressão Logística, os maiores p-valores apresentados na base de treino foram retirados um a um. Dada à capacidade computacional disponível e o tamanho do banco de dados, o teste de Stepwiser não foi utilizado por não realizar simultaneamente todas as interações necessárias nas observações. O critério do p-valor foi escolhido em função de sua eficiência, simplicidade e utilização de menor capacidade computacional.

A variável curso contem 408 grupos totalizando 407 variáveis *dummy*, o que torna inviável seu uso neste modelo. A variável denominada "PNE" foi retirada por não ter estudantes deficientes com representatividade suficiente para dizer se isso implica em inadimplência.

As variáveis que não entraram no modelo, de acordo com o critério do p-valor, foram a área de conhecimento do curso, estado civil, sexo e idade.

A Tabela 57 mostra um resumo do modelo de Regressão Logística com a variável, a categoria, estimativa do intercepto β_0, o erro padrão estimado, o valor da estimativa Z e o p-valor.

As categorias foram nomeadas como A, B, C e D, de acordo com a quantidade de faixa de valor, para um melhor entendimento. Como para as variáveis *dummy* sempre aparecem *n-1* categorias, as primeiras, denominadas de categorias A, foram retiradas. A escolha dessa faixa de categoria foi feita por padrão do *software* Rstudio. Foi utilizado um truncamento depois da terceira casa decimal das variáveis da estimativa e do erro padrão, para uma melhor organização.

Tabela 34 – Resumo do modelo de Regressão Logística utilizado para classificar os estudantes inadimplentes que tem financiamento estudantil utilizando os egressos do Fies até setembro de 2017

Variável	Categoria	Estimativa	Erro padrão	Valor Z	P-valor
Intercepto	-	-1,626628	0,044164	-36,832	$< 2e^{-16}$
Quantidade de semestres contratados	B	0,171147	0,013985	12,238	$< 2e^{-16}$

(continua)

Tabela 34 – Resumo do modelo de Regressão Logística utilizado para classificar os estudantes inadimplentes que tem financiamento estudantil utilizando os egressos do Fies até setembro de 2017 (continuação)

Variável	Categoria	Estimativa	Erro padrão	Valor Z	P-valor
UF da IES	B	0,162817	0,010826	15,040	$< 2e^{-16}$
UF da IES	C	-0,570688	0,015114	-37,760	$< 2e^{-16}$
Garantia	B	0,393168	0,043232	9,094	$< 2e^{-16}$
Garantia	C	0,465600	0,039787	11,702	$< 2e^{-16}$
Valor financiado	B	-0,705960	0,009426	-74,898	$< 2e^{-16}$
Valor financiado	C	-1,307815	0,013116	-99,715	$< 2e^{-16}$
Turno	B	0,070797	0,009829	7,203	$5.90e^{-13}$
Taxa de juros	B	-1,420991	0,061867	-22,968	$< 2e^{-16}$
Percentual financiado	B	-0,372896	0,011060	-33,715	$< 2e^{-16}$
Valor da renda dos fiadores	B	-0,105011	0,020793	-5,050	$4.41e^{-07}$
Valor da renda dos fiadores	C	-0,453075	0,020049	-22,599	$< 2e^{-16}$
Valor da renda dos fiadores	D	1,480102	0,023649	62,585	$< 2e^{-16}$
Renda pessoal mensal bruta	B	0,067052	0,010227	6,556	$5.52e^{-11}$
Renda pessoal mensal bruta	C	-0,105847	0,009742	-10,865	$< 2e^{-16}$
Renda familiar mensal bruta	B	-0,236550	0,009038	-26,174	< 2e-16
Renda familiar mensal bruta	C	-0,508071	0,012753	-39,841	< 2e-16
Situação da inscrição	Encerrado	-0,908683	0,013859	-65,568	< 2e-16
Prouni	Sim	-0,568701	0,020777	-27,371	< 2e-16
Ensino médio	Pública	0,166104	0,012768	13,009	< 2e-16
Banco	Caixa	0,008376	0,008376	34,417	< 2e-16

Fonte: os autores (2018).

Pelo sinal do intercepto β, na categoria estimativa e pela sua grandeza pode-se ter uma boa ideia do perfil do estudante inadimplente. Ou seja, valores positivos indicam característica de estudantes mais inadimplentes e valores negativos características de adimplência. As únicas variáveis que não tiveram p-valor ótimo ($2e^{-16}$) foram turno, valor da renda dos fiadores (categoria B) e renda pessoal mensal bruta (categoria B), porém essas variáveis apresentaram um p-valor com significância maior que 95%.

A importância de cada variável que influencia na inadimplência e por isso permaneceu no modelo está relacionada na Tabela 35.

Tabela 35 – Percentual de importância de cada variável que influencia na inadimplência de acordo com a Regressão Logística que foi utilizado para classificar os estudantes inadimplentes que tem financiamento estudantil utilizando os egressos do Fies até setembro de 2017

Variável	Categoria	Overall (%)
Quantidade de semestres contratados	B	12,237676
Região da IES	B	15,039855
Região da IES	C	37,759902
Garantia	B	9,094299
Garantia	C	11,702352
Valor financiado	B	74,897903
Valor financiado	C	99,714728
Turno	B	7,202803
Taxa de juros	B	22,968396
Percentual financiado	B	33,715065
Valor da renda dos fiadores	B	5,050206
Valor da renda dos fiadores	C	22,598536
Valor da renda dos fiadores	D	62,585327
Renda pessoal mensal bruta	B	6,556122
Renda pessoal mensal bruta	C	10,864660
Renda familiar mensal bruta	B	26,173738
Renda familiar mensal bruta	C	39,840801
Situação da inscrição	Encerrado	65,567797
Prouni	Sim	27,371015
Ensino médio	Pública	13,009311
Banco	Caixa	34,416674

Fonte: os autores (2018).

A Tabela 35 apresenta a porcentagem de importância de cada variável do modelo em uma escala de 0 a 100% para cada categoria. A título de exemplo observa-se que o valor financiado (categoria C) é 99,71% relevante, sendo, pois, a variável com maior percentual de importância. Essa interpretação corrobora com o β negativo apresentado na Tabela 57, ou seja, estudantes com os maiores valores financiados tendem a ser mais inadimplentes. Tais interpretações podem ser ampliadas utilizando-se todas as variáveis da tabela em referencia.

7.2 RANDOM FOREST

A mesma base de teste e de treino que foi utilizada para a Regressão Logística foi utilizada para o Random Forest.

Essa técnica apresenta um ranking da importância das variáveis, por isso não é necessário retirá-las. Por ser uma técnica de aprendizagem de máquina é interessante que ela não fique "engessada" sempre precisando de alguém para controlar, por esse motivo foi preferível retirar a categorização.

Como essa técnica usa árvores de decisões combinadas ao invés de variáveis *dummy*, as variáveis curso e estados da IES podem ser acrescentados e as variáveis áreas de conhecimento e região continuaram, pois esse modelo não apresenta pressuposto da necessidade de não correlação.

A Tabela 36 apresenta também as seguintes colunas: importância relativa, escala de importância e porcentagem de importância de todas as variáveis em escala decrescente de prioridade para o cálculo da inadimplência.

A Tabela 36 apresentada pelo modelo Random Forest:

Tabela 36 – Resumo do Random Forest utilizado para classificar os estudantes inadimplentes que tem financiamento estudantil utilizando os egressos do Fies até setembro de 2017

Variáveis	Importância relativa	Escala de importância	Porcentagem de importância
Valor da renda dos fiadores	205.348,172	1,00000000	0,166496766
Curso	196.848,266	0,95860734	0,159605022
Valor financiado	123.524,688	0,60153780	0,100154098
Estado da IES	110.587,273	0,53853547	0,089664413
Renda familiar mensal bruta	106.471,961	0,51849481	0,086327708

(continua)

Tabela 36 - Resumo do Random Forest utilizado para classificar os estudantes inadimplentes que tem financiamento estudantil utilizando os egressos do Fies até setembro de 2017 (continuação)

Variáveis	Importância relativa	Escala de importância	Porcentagem de importância
Tipo de garantia	105.853,039	0,51548080	0,085825885
Idade	65.609,234	0,31950240	0,053196117
Quantidade de semestres contratados	55.977,805	0,27259948	0,045386932
Renda pessoal mensal bruta	47.427,848	0,23096309	0,038454607
Percentual financiado	39.055,254	0,19019041	0,031666089
Região da IES	31.416,961	0,15299362	0,025472944
Estado civil	26.852,842	0,13076738	0,021772345
Situação da inscrição	21.888,359	0,10659145	0,017747132
Área de conhecimento	21.677,148	0,10556290	0,017575881
Turno	20.473,012	0,09969902	0,016599565
Banco	15.816,619	0,07702342	0,012824151
Sexo	14.334,732	0,06980696	0,011622634
Ensino médio	11.866,350	0,05778649	0,009621263
Prouni	6.209,148	0,03023717	0,005034392
Taxa de juros	4.001,296	0,01948542	0,003244260
PNE	2.106,306	0,01025724	0,001707798

Fonte: os autores (2018).

A Tabela 36 apresenta, entre outros, valor da renda dos fiadores, curso e valor financiado. Essas são as três variáveis mais importantes para explicar a inadimplência, em conformidade com o modelo Random Forest.

8 COMPARAÇÃO DOS MODELOS

8.1 EFICIÊNCIA DO MODELO DE REGRESSÃO LOGÍSTICA

O modelo foi aplicado à base teste, o ponto de corte escolhido foi o do score igual a 0,3. Esse valor foi escolhido observando o ponto em que a acurácia era maior e que acertava a maior quantidade de inadimplentes, esse ponto foi escolhido observando essas duas variáveis, pois não adiantaria o modelo classificar todos como bons pagadores tendo uma acurácia muito alta e não dividindo corretamente os alunos. A matriz de confusão encontrada foi a seguinte:

Tabela 37– Matriz de confusão dos valores referente a Regressão Logística em relação estudantes inadimplentes que tem financiamento estudantil utilizando os egressos do Fies até setembro de 2017 e os valores preditos

		Valores de referencia		
		Adimplente	**Inadimplente**	**Total**
Valores preditos	**Adimplente**	94.171	12.144	106.315
	Inadimplente	39.473	32.376	71.849
	Total	133.644	44.520	178.164

Fonte: os autores (2018).

A Tabela 37 mostra que o modelo, de Regressão Logística, classificou 94.171 estudantes que eram adimplentes e 32.376 estudantes inadimplentes nas categorias corretas. E 39.473 estudantes que eram adimplentes como inadimplentes e, além de 12.144 inadimplentes como adimplentes.

Observe as proporções da Tabela 38 com um truncamento depois da segunda casa decimal:

Tabela 38 – Matriz de confusão das proporções referentes à Regressão Logística em relação estudantes inadimplentes que tem financiamento estudantil utilizando os egressos do Fies até setembro de 2017 e os valores preditos

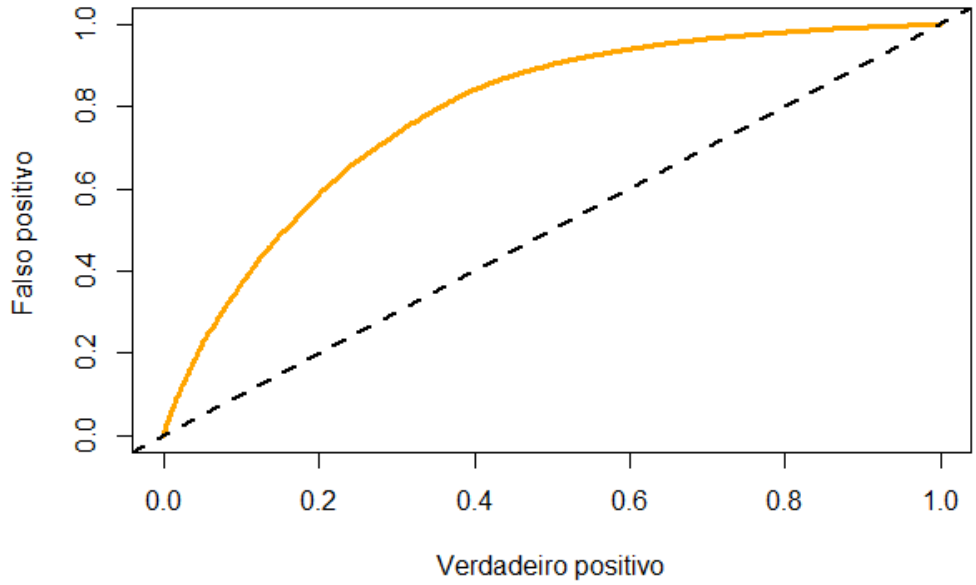
		Proporções de referencia (%)		
		Adimplente	**Inadimplente**	**Total**
Proporções preditas (%)	**Adimplente**	53,51	7,04	60,58
	Inadimplente	21,51	17,91	39,52
	Total	75,04	24,96	100

Fonte: os autores (2018).

A seguir estão mais algumas medidas para a validação do modelo:

- A acurácia foi de 0,70 mostrando que o modelo apresenta 70% de acerto;
- A medida de Kappa foi de 0.36 com um p-valor dede $2{,}2e^{-16}$ e de acordo com essa media essa medida o classificar de regressão logístico é fraco;
- A sensibilidade desse modelo foi 0,73 mostrando que o modelo acerta 73% dos adimplentes;
- A especificidade foi de 0,70 mostrando que esse modelo acerta 70% dos inadimplentes;
- A curva ROC segue abaixo com o a área igual a 0,77 e quanto maior a área melhor o modelo, sendo um bom indicador para comparar dois modelos.

Gráfico 1 – Curva ROC do modelo de Regressão Logística



Fonte: os autores (2018).

8.2 EFICIÊNCIA DO MODELO RANDOM FOREST

O Random Forest apresenta a matriz de confusão calculada com a própria base de treino, contudo, para se evitar que a matriz de confusão fique viesada, ela foi calculada utilizando-se a base de teste para obtenção de uma melhor validação do modelo.

Tabela 39 – Matriz de confusão dos valores referente a Random Forest em relação estudantes inadimplentes que tem financiamento estudantil utilizando os egressos do Fies até setembro de 2017 e os valores preditos

		Valores de referencia		
		Adimplente	**Inadimplente**	**Total**
Valores preditos	**Adimplente**	114.607	11.355	125.962
	Inadimplente	19.037	33.165	52.202
	Total	133.644	44.520	178.164

Fonte: os autores (2018).

A Tabela 39 mostra que do total de estudantes da base de teste o modelo classificou 114.607 estudantes adimplentes realmente como adimplentes e 33.165 inadimplentes na categoria correta. Ela também classificou 11.355 que eram inadimplentes como adimplentes e 19.037 adimplentes como inadimplentes.

A seguir temos a proporção dos valores da matriz de confusão com truncamento depois da segunda casa decimal:

Tabela 40 – Matriz de confusão das proporções referentes à Random Forest em relação estudantes inadimplentes que tem financiamento estudantil utilizando os egressos do Fies até setembro de 2017 e os valores preditos

		Proporções de referencia (%)		
		Adimplente	**Inadimplente**	**Total**
Proporções preditas (%)	**Adimplente**	64,30	6,40	70,70
	Inadimplente	10,70	18,60	25,0
	Total	75,00	25,00	100

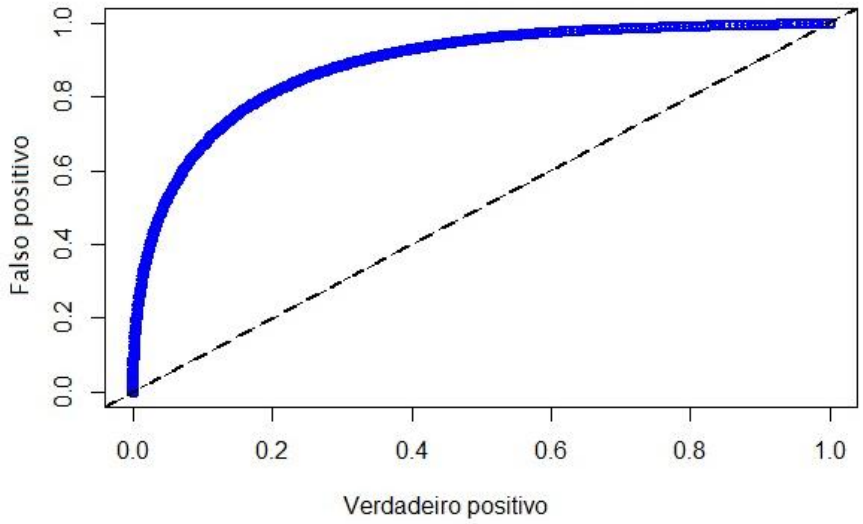
Fonte: os autores (2018).

A seguir estão mais algumas medidas para a validação do modelo:

- A acurácia foi de 0,82 mostrando que o modelo apresenta 82% de acerto;
- A medida de Kappa foi de 0.56 com um p-valor dede $2{,}2e^{-16}$ e de acordo com essa media essa medida o classificar de regressão logístico é razoável;
- A sensibilidade desse modelo foi 0,86 mostrando que o modelo acerta 86% dos adimplentes;
- A especificidade foi de 0,73 mostrando que esse modelo acerta 73% dos inadimplentes;

– A curva ROC segue abaixo com o a área igual a 0,86 e quanto maior a área melhor o modelo, sendo um bom indicador para comparar dois modelos.

Gráfico 2 – Curva ROC do modelo Random Forest



Fonte: os autores (2018).

De acordo com os dados apresentados o modelo Random Forest se mostrou mais eficiente que o modelo de Regressão logística para esse banco de dados.

9 APLICAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA

O modelo prevê que entre os estudantes que, em setembro de 2017, estão nas fases de utilização e carência 66,89% vão ser adimplentes e 33,11% vão ser inadimplentes quando chegarem à fase de amortização.

Tabela 41 – Percentual de adimplência e inadimplência prevista para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

Adimplentes (%)	Inadimplentes (%)
66,89	33,11

Fonte: os autores (2018).

O montante esperado de arrecadação é de R$ 65.191.140.957,00. Por esse modelo, espera-se um déficit de arrecadação de R$ 21.584.786.770,86 no que se refere ao valor financiado aos estudantes indicados como maus pagadores.

9.1 APLICAÇÃO DA REGRESSÃO LOGÍSTICA DIVIDIDA EM FASE

Tabela 42 – Porcentagem de previsão de adimplência e inadimplência dividida entre as fases para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017 de acordo com a Regressão Logística

Fase	Adimplente (%)	Inadimplente (%)
Utilização	64,54	35,45
Carência	52,78	47,21

Fonte: os autores (2018).

A quantidade estimada de estudantes inadimplentes com o contrato que em Setembro de 2017 estavam na fase de carência é de 47,21% o que é maior do que a quantidade estimada de inadimplentes que estavam na fase de utilização.

9.2 APLICAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA POR REGIÃO

A seguir temos a estimativa de inadimplência por região para o modelo de Regressão Logística.

Tabela 43 – Porcentagem de adimplência e inadimplência prevista por região para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017 de acordo com a Regressão Logística

Região	Adimplente (%)	Inadimplente (%)
Norte	54,66	45,33
Nordeste	57,13	42,86
Sudeste	61,00	38,99
Sul	94,89	5,10
Centro-Oeste	42,48	57,51

9.3 APLICAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA POR ÁREA DE CONHECIMENTO

A seguir estão as previsões de adimplência e inadimplência, divididas em áreas de conhecimento.

Tabela 44 – Porcentagem de adimplência e inadimplência prevista por área de conhecimento para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017 de acordo com a Regressão Logística

Área de conhecimento	Adimplente (%)	Inadimplente (%)
Educação	45,22	54,77
Humanidades e artes	58.78	41.21
Ciências Sociais, Negócios e Direito	58.97	41.02
Ciências, Matemática e Computação.	60.75	39.24
Engenharia, Produção e Construção	67.70	32.29
Agricultura e Veterinária	74.12	25.87
Saúde e Bem estar Social	63.28	36.71
Serviços	56.97	43.02

Fonte: os autores (2018).

9.4 APLICAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA POR ÁREA DE RISCO DO BANCO.

A Tabela 45 comparação do risco que o Bacen indica e que o modelo de Regressão Logística mostra para esses estudantes.

Tabela 45 – Comparação do risco do modelo logístico com o risco sugerido pelo Bacen para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

Bacen		Modelo logístico	
Faixa de risco	Inadimplência (%)	Adimplência (%)	Inadimplência (%)
A	0,5	71,58	28,41
B	1	62,79	37,23
C	3	60,58	39,41
D	10	55,87	44,12
E	30	53,51	46,48
F	50	36,95	63,04
G	70	33,33	66,66
H	100	26,72	73,27

Fonte: os autores (2018).

A Tabela 45 mostra o percentual de provisionamento que o Bacen indica de forma genérica e que o percentual provisionado de acordo com o modelo de Regressão Logística. Percebe-se que o provisionamento indicado pelo Bacen espera que nenhum estudante volte a pagar depois dos 356 dias e o modelo de Regressão Logística admite essa hipótese.

9.5 PERDA ESPERADA DOS ESTUDANTES EM UTILIZAÇÃO E CARENCIA

Tabela 46 – Comparação do provisionamento do modelo de Regressão Logística com o sugerido pelo Bacen para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

	Bacen		**Regressão logística**	
Faixa de risco	**Inad. Esperada (%)**	**Provisionamento (R$)**	**Inad. Esperada (%)**	**Provisionamento (R$)**
A	0,5	R$ 221.155.942,00	28,41	R$12.566.080.605,00
B	1	R$118.305.007,00	37,23	R$4.404.495.416,00
C	3	R$133.736,90	39,41	R$1.756.857,00
D	10	R$778.700,00	44,12	R$ 3.435.624,00
E	30	R$951.464.057,00	46,48	R$1.474.134.978,00
F	50	R$239.017,40	63,04	R$301.353,10
G	70	R$133.698,70	66,66	R$127.319,40
H	100	R$5.944.991.125,00	73,27	R$4.355.894.997,00
Total	11,44	R$ 7.237.201.284,00	34,98	**R$ 22.806.227.149,50**

Fonte: os autores (2018).

Utilizando a mesma forma de calcular do BACEN, considerando que todos os estudantes devem a mesma coisa e multiplicando a dívida pela porcentagem que se espera de inadimplência o modelo de Regressão Logística apresenta um total de R$ 22.806.227.149,50 de dívida esperada, para os estudantes que ainda não começaram a pagar, enquanto o Bacen aponta um total de R$ 7.237.201.284,00 de déficit orçamentário para esses mesmos estudantes.

9.6 PERDA ESPERADA DOS ESTUDANTES EM TODAS AS FAZES

Para esse trabalho foi utilizado o mesmo mês utilizado para a construção do modelo para esse cálculo. O que pode gerar certo viés, porém esses dados são um bom indício da aplicação para outros meses.

Tabela 47 – Comparação do provisionamento do modelo de Regressão Logística com o sugerido pelo Bacen para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

	Bacen		**Regressão logística**	
Faixa de risco	**Inad. Esperada (%)**	**Provisionamento (R$)**	**Inad. Esperada (%)**	**Provisionamento (R$)**
A	0,5	R$ 260.783.142,00	28,41	R$14.817.698.129,00
B	1	R$ 128.250.258,00	37,23	R$ 4.774.757.093,00
C	3	R$ 785.2862,00	39,41	R$ 103.160.430,00
D	10	R$ 2.799.9624,00	44,12	R$ 123.534.342,00
E	30	R$1.053.009.774,00	46,48	R$ 1.631.463.143,00
F	50	R$ 33.790.719,00	63,04	R$ 42.603.338,00
G	70	R$ 49.747.420,00	66,66	R$ 47.373.758,00
H	100	R$ 8.883.476.612,00	73,27	R$ 6.508.923.314,00
Total	11,44	R$ 10.444.910.411,00	34,98	**R$ 28.049.513.547,00**

Fonte: os autores (2018).

Calculando a dívida esperada do mesmo modo que o BACEN, considerando que todos os estudantes devem a mesma coisa e multiplicando a dívida pela porcentagem que se espera de inadimplência encontra-se um total de R$ 10.444.910.411,00 de déficit orçamentário enquanto o modelo de Regressão Logística apresenta um total de R$ 28.049.513.547,00 de dívida esperada.

10 APLICAÇÃO DO RANDOM FOREST

O modelo prevê que entre os estudantes que, em setembro de 2017, estão nas fases de utilização e carência 71,88% serão adimplentes e 28,11% inadimplentes quando chegarem à fase de amortização.

Tabela 48 – Percentual de adimplência e inadimplência prevista de acordo com o modelo Random Forest para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

Adimplentes (%)	Inadimplentes (%)
71,88	28,11

Fonte: os autores (2018).

Do total de R$ 65.191.140.957,00 que se espera que os estudantes paguem, do valor referente aos cursos, esse modelo prevê não receber R$18.325.229.723,00, sendo esse valor a soma dos cursos que o Random Forest indicou como mau pagadores.

10.1 APLICAÇÃO DO RANDOM FOREST DIVIDIDO EM FASE

Tabela 49 – Porcentagem de previsão de adimplência e inadimplência dividida entre as fases que ainda não começam a pagar de acordo com o Random Forest para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

Fase	Adimplente (%)	Inadimplente (%)
Utilização	73,18	26,18
Carência	68,27	31,72

Fonte: os autores (2018).

Mesmo que a mudança não seja tão expressiva e a quantidade de inadimplência continue alta pode-se perceber que a quantidade de inadimplência em relação aos estudantes em carência é maior do que com os estudantes que estão em utilização.

10.2 APLICAÇÃO DO MODELO RANDOM FOREST POR REGIÃO

Na tabela 50 estão às proporções das previsões de inadimplência por região, pelo modelo Random Forest.

Tabela 50 – Porcentagem de adimplência e inadimplência prevista por região de acordo com o modelo Random Forest para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

Região	Adimplente (%)	Inadimplente (%)
Norte	58.96	41.03
Nordeste	65.05	34.94
Sudeste	73.93	26.06
Sul	93.37	6.62
Centro-oeste	65.92	34.07

Fonte: os autores (2018).

Percebe-se que região Sul com uma estimativa de 93,37% de adimplência destoa das outras regiões mostrando a necessidade de se entender o porquê dessa região terem estudantes que são melhores pagadores.

10.3 APLICAÇÃO DO MODELO RANDOM FOREST POR ÁREA DE CONHECIMENTO

A Tabela 51 apresenta em formato de tabelas as proporções das previsões de inadimplência, em áreas de conhecimento de acordo com o manual do CNPq, pelo modelo Random Forest.

Tabela 51 – Porcentagem de adimplência e inadimplência prevista por área de conhecimento de acordo com o Random Forest para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

Área de conhecimento	Adimplente (%)	Inadimplente (%)
Educação	44.48	55.51
Humanidades e artes	59,03	40,96
Ciências Sociais, Negócios e Direito	69.80	30.19
Ciências, Matemática e Computação	73.59	26.40
Engenharia, Produção e Construção	84.16	15.83
Agricultura e Veterinária	89.91	10.08
Saúde e Bem estar Social	70.75	29.24
Serviços	53.31	46.68

Fonte: os autores (2018).

10.4 APLICAÇÃO DO MODELO RANDOM FOREST POR CURSO

Para a construção da tabela que mostra a inadimplência esperada dividida por curso só foi possível utilizando o modelo Randon Forest, pois é uma variável qualitativa com muitas observações, o que torna inviável categorizar os cursos. E mesmo utilizar essa variável sem categorizar seria inviável pela quantidade de variáveis *dummy*

Em alguns cursos apenas um aluno é beneficiário do programa, e com base em números muito pequenos de amostras não é possível estimar uma porcentagem de inadimplência. Por esse motivo foi decidido fazer uma tabela com as estimativas de inadimplência apenas dos 25 cursos mais financiados no semestre de referencia. A Tabela 75 se encontra em ordem decrescente de quantidade de alunos que utilizaram o fies para pagar o respectivo curso, em setembro de 2017.

Tabela 52 – Estimativa de inadimplência dos 25 cursos mais financiados no segundo semestre de 2017 de acordo com o modelo Random Forest para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

Ranking de financiamento	Curso	Adimplente (%)	Inadimplente (%)
01	Direito	75,59	24,40
02	Engenharia civil	85,30	14,69
03	Enfermagem	66,62	33,37
04	Administração	60,14	39,85
05	Psicologia	77,97	22,02
06	Fisioterapia	65,38	34,61
07	Ciências contábeis	70,55	29,44
08	Educação física	46,57	53,42
09	Pedagogia	42,27	57,72
10	Arquitetura e urbanismo	86,01	13,98
11	Engenharia de produção	78,28	21,71
12	Engenharia mecânica	86,88	13,11
13	Odontologia	92,41	7,58
14	Nutrição	65,49	34,50
15	Farmácia	83,55	16,44
16	Medicina	99,83	0,16
17	Engenharia elétrica	83,05	16,94

(continua)

Tabela 52 – Estimativa de inadimplência dos 25 cursos mais financiados no segundo semestre de 2017 de acordo com o modelo Random Forest para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017 (continuação)

Ranking de financiamento	Curso	Adimplente (%)	Inadimplente (%)
18	Medicina veterinária	91,49	8,50
19	Biomedicina	73,72	26,27
20	Serviço social	40,32	59,67
21	Comunicação Social	62,68	37,31
22	Agronomia	88,01	11,98
23	Sistemas de informação	74,11	25,88
24	Gestão de recursos humanos	24,05	75,94
25	Engenharia de controle e automação	83,51	16,48

Fonte: os autores (2018).

10.5 APLICAÇÃO DO MODELO DE RANDOM FOREST POR ÁREA DE RISCO DO BANCO

A seguir temos a Tabela 53 de comparação do risco que o Bacen indica e que o Random Forest mostra para esses estudantes.

Tabela 53 – Comparação do risco do modelo Random Forest com o risco sugerido pelo Bacen para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

Bacen		Modelo Random Forest	
Faixa de risco	Inadimplência (%)	Adimplência (%)	Inadimplência (%)
A	0,5	81,25	18,74
B	1	75,61	24,38
C	3	63,07	36,92
D	10	65,12	34,87
E	30	63,69	36,30
F	50	28,26	71,73
G	70	38,09	61,90
H	100	37,22	62,77

Fonte: os autores (2018).

A Tabela 53 mostra que o percentual esperado de inadimplência do Modelo Random Forest é diferente do provisionado pelo Bacen e o Modelo Random Forest admite que o estudante volte a pagar depois de decorridos 365 dias que o atraso começou.

10.6 PERDA ESPERADA DOS ESTUDANTES EM UTILIZAÇÃO E CARENCIA

Tabela 54 – Comparação do provisionamento do modelo Random Forest com o sugerido pelo Bacen para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

	Bacen		**Random Forest**	
Faixa de risco	**Inad. Esperada (%)**	**Provisionamento (R$)**	**Inad. Esperada (%)**	**Provisionamento (R$)**
A	0,5	R$ 221.155.942,00	18,74	R$8.288.924.694,00
B	1	R$118.305.007,00	24,38	R$2.884.276.074,00
C	3	R$133.736,90	36,92	R$1.645.856,00
D	10	R$778.700,00	34,87	R$2.716.884,00
E	30	R$951.464.057,00	36,30	R$1.151.271.508,00
F	50	R$239.017,40	71,73	R$ 342.894,30
G	70	R$133.698,70	61,90	R$118.227,90
H	100	R$5.944.991.125,00	62,77	R$3.731.670.929,00
Total	11,44	R$ 7.237.201.284,00	24,64	**R$ 16.060.967.067,20**

Fonte: os autores (2018).

Calculando a dívida esperada do mesmo modo que o BACEN, considerando que todos os estudantes devem a mesma coisa e multiplicando a dívida pela porcentagem que se espera de inadimplência encontra-se um total de R$ 7.237.201.284,00 de déficit orçamentário enquanto o modelo de Random Forest apresenta um total de R$ 16.060.967.067,20 de dívida esperada.

10.7 PERDA ESPERADA DOS ESTUDANTES EM TODAS AS FAZES

Calcularam-se os resultados mostrados na Tabela 55 com o mesmo banco de dados utilizado para se obter o modelo, o que pode gerar certo viés, porém esse resultado exemplifica uma aplicação que pode ser realizada com os dados referentes a outros meses.

Tabela 55 – Comparação do provisionamento do modelo Random Forest com o sugerido pelo Bacen para os estudantes que estão nas fases de utilização e carência do contrato com o Fies em Setembro de 2017

	Bacen		**Random Forest**	
Faixa de risco	**Inad. Esperada (%)**	**Provisionamento (R$)**	**Inad. Esperada (%)**	**Provisionamento (R$)**
A	0,5	R$ 260.783.142,00	18,74	R$ 9.774.152.163,00
B	1	R$ 128.250.258,00	24,38	R$ 3.126.741.282,00
C	3	R$ 785.2862,00	36,92	R$ 96.642.554,00
D	10	R$ 2.799.9624,00	34,87	R$ 97.690.689,00
E	30	R$1.053.009.774,00	36,30	R$ 1.274.141.827,00
F	50	R$ 33.790.719,00	71,73	R$ 48.476.165,00
G	70	R$ 49.747.420,00	61,90	R$ 43.990.933,00
H	100	R$ 8.883.476.612,00	62,77	R$ 5.576.158.269,00
Total	11,44	R$ 10.444.910.411,00	24,64	**R$ 20.037.993.882,00**

Fonte: os autores (2018).

Calculando a dívida esperada do mesmo modo que o BACEN, considerando que todos os estudantes devem a mesma coisa e multiplicando a dívida pela porcentagem que se espera de inadimplência encontra-se um total de R$ 10.444.910.411,00de déficit orçamentário enquanto o modelo de Random Forest apresenta um total de **R$ 20.037.993.882,00** de dívida esperada.

11 CONCLUSÃO

O financiamento estudantil pode ser um uma excelente forma de o governo melhorar o índice de escolarização da população, ampliando as possibilidades de quem não tem condições de pagar e/ou não conseguiu vaga em uma universidade pública consiga um melhor índice de escolarização.

Apresentou-se uma análise do que se espera dos estudantes que participaram do Fies. Uma análise descritiva mostra que a maioria dos estudantes obteve o financiamento com juros a 3,4% ao ano, o que pode ser prejudicial para o governo visto que estudantes que teriam condições de arcar com os custos da universidade podem preferir investir o dinheiro e pagar com financiamento depois do grande prazo de amortização, assim obtendo lucros. Estudantes que obtiveram um financiamento com juros mais elevados tendem a ser melhores pagadores e levanta-se a hipótese de apesar desses estudantes pagarem um valor maior serem estudantes que realmente necessitavam do financiamento. Essa investigação ficou como uma sugestão pera outro estudo.

Os estudantes da região Sul do país também mostraram ter um índice de inadimplência menor comparado a outras regiões e levanta-se a hipótese de que as faculdades dessa região se não tiverem um ensino melhor, proporcionando melhores oportunidades no mercado de trabalho, ajuda a formar o caráter das pessoas.

A idade média dos estudantes que aderiram o financiamento foi de 23 anos, mostrando que provavelmente a maior parte dos estudantes que procuram o financiamento estudantil o faz logo após a conclusão do ensino médio.

O Fgduc e a Fiança convencional agrupam quase a totalidade dos estudantes, seguido pela combinação dessas duas modalidades de financiamento e sugere-se investigar o que tem ocorrido com os outros tipos de financiamento.

Percebe-se que apesar de existirem rendas familiares e renda dos fiadores muito elevadas essa situação não acontece na grade maioria dos casos.

O Random Forest se mostrou um modelo mais eficiente para esses dados, segundo todas as meditas utilizadas para comparar, do que o modelo de Regressão Logística. Essa maior eficiência já era esperada, pois foi possível acrescentar duas variáveis a mais: “curso” e “estado da IES” e essas variáveis se mostraram muito importantes para entender a inadimplência. Esse modelo também aponta uma quantidade de inadimplentes e um valor de dívida não liquidada menor e por isso sugere-se a utilização desse modelo para as previsões futuras.

O cálculo de provisionamento do Bacen do valor que se espera não ser recebido mostrou-se longe da realidade observada nos dois modelos. Essa diferença já era esperada, pois o Randon Forest e o Modelo de Regressão Logística foram construídos especificamente para os dados de financiamento do Fies e o Banco utiliza a mesma forma de provisionamento para diferentes bancos de dados em diferentes situações.

O Bacen faz essa análise de uma forma mais empírica, utilizando a mesma forma de provisionamento para diferentes situações e banco de dados.

Para qualquer forma de modelo adotada como provisionamento espera-se uma grande quantidade de estudantes que não honrem com a dívida adquirida somando um alto valor para ser possível que o Fgduc honre com a dívida desses estudantes inadimplentes e garanta a continuidade do programa.

REFERÊNCIAS

ALTMAN, Edward I.; SAUNDERS, Anthony. Credit risk measurement: developments over the last 20 years. **Journal of Banking and Finance**, Virginia, v. 21, n. 11/12, p. 1721-1742, 1997.

ANDRADE, Simone Ferreira Capriccio et al. A inadimplência nas instituições particulares de ensino na cidade de Franca. **FACEF Pesquisa-Desenvolvimento e Gestão**, Franca, v. 11, n. 1, p. 45-58, 2010.

BANCO CENTRAL DO BRASIL (BACEN). **Resolução nº 2682**. Dispõe sobre critérios de classificação das operações de crédito e regras para constituição de provisão para créditos de liquidação duvidosa. Brasília, DF, 22 dez. 1999. Disponível em: <http://www.bcb.gov.br/pre/normativos/res/1999/pdf/res_2682_v2_L.pdf>. Acesso em: 15 ago. 2017.

BRASIL. Lei n° 10.260, de 12 de julho de 2001. Dispõe sobre o Fundo de Financiamento ao estudante do Ensino Superior e dá outras providências. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 13 jul. 2001. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/LEIS_2001/L10260.htm>. Acesso em: 16 ago. 2017.

BREIMAN, Leo. Random forests. **Machine Learning**, Cambridge, v. 45, n. 1, p. 5-32, 2001.

CHAPMAN, Bruce. Income contingent loans for higher education: international reform. In: HANUSHEK, Eric A.; WELCH, Finis. **Handbook of the economics of education**. Haarlem: Elsevier, 2006. v. 1, p. 1435-1503.

DE PADRÕES, Reconhecimento; ASCENSO, João; FRED, Ana. Reconhecimento de Padrões.

DOMINGOS, Pedro. A few useful things to know about machine learning. **Communications of the ACM**, New York, v. 55, n. 10, p. 78-87, 2012.

DOWD, Alicia C.; COURY, Tarek. The effect of loans on the persistence and attainment of community college students. **Research in higher education**, [New York], v. 47, n. 1, p. 33-62, 2006.

DYNARSKI, Susan. An economist's perspective on student loans. In: CESIFO AREA CONFERENCE ON THE ECONOMICS OF EDUCATION, 2014, Munich. **Annals…** Munich: CESifo, 2014.

ENGELMANN, Bernd; HAYDEN, Evelyn; TASCHE, Dirk. Testing rating accuracy. **Risk**, [S.l.], v. 16, n. 1, p. 82-86, 2003.

GERTLER, Mark; KARADI, Peter. Monetary policy surprises, credit costs, and economic activity. **American Economic Journal: Macroeconomics**, Nashville, v. 7, n. 1, p. 44-76, 2015.

HOSMER, David W.; LEMESHOW, Stanley. **Applied regression analysis**. New York, John Willey, 1989.

LANDIS, J. Richard; KOCH, Gary G. The measurement of observer agreement for categorical data. **Biometrics**, Washington, v. 33, n. 1, p. 159-174, 1977.

LANTZ, Brett. **Machine learning with R**. Birmingam: Packt Publishing, 2015.

NASCIMENTO, Paulo A. Meyer M.; LONGO, Gustavo Frederico. Qual o peso dos encargos de reembolso sobre a renda esperada dos beneficiários do FIES?. **Boletim Radar**, Brasília, DF, n. 46, p. 23-43, 2016.

PARKINSON, Kenneth L.; OCHS, Joyce R. Using credit screening to manage credit risk. **Business Credit**, [S.l.], v. 22, p. 23-27, 1998.

PIRAMUTHU, Selwyn. On preprocessing data for financial credit risk evaluation. **Expert Systems with Applications**, London, v. 30, n. 3, p. 489-497, 2006.

PRESS, S. James; WILSON, Sandra. Choosing between logistic regression and discriminant analysis. **Journal of the American Statistical Association**, Alexandria, v. 73, n. 364, p. 699-705, 1978.

ROCHA, Wilsimara M.; EHRL, Philipp; MONASTERIO, Leonardo M. Análise de impacto do FIES sobre o salário do trabalhador formal. In: ENCONTRO NACIONAL DE ECONOMIA, 44., 2016, Foz do Iguaçu. **Anais...** Niterói: ANPEC, 2016.

ROCHA, Wilsimara M.; MONASTERIO, Leonardo M.; EHRL, Philipp. Qual foi o impacto do FIES nos salários?. **Boletim Radar**, Brasília, DF, n. 46, p. 33-38, 2016.

SALMI, Jamil. The challenge of sustaining student loan system: Colombia and Chile. **International Higher Education**, Chestnut Hill, n. 72, p. 21-23, 2013.

SAUNDERS, Anthony. **Medindo o risco de crédito**: novas abordagens para value at risk e outros paradigmas. Rio de Janeiro: Qualitymark, 2000.

ST. JOHN, Edward. P. et al. Economic influences on persistence reconsidered: how can finance research inform the reconceptualization of persistence models? In: BRAXTON, John M. (Ed.). **Reworking the student departure puzzle**. Nashville: Vanderbilt University Press, 2000.

TRIBUNAL DE CONTAS DA UNIÃO (TCU). **Relatório de Auditoria TC 011.884/2016-9**. Brasília, DF, 2016.

VICENTE, Ernesto F. R. **A estimativa do risco na constituição da PDD**. 2001. 179 f. Dissertação (Mestrado em Controladoria e Contabilidade) – Faculdade de Economia Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2001.

VIEIRA, José Rômulo de Castro. **Predição do bom e do mau pagador no programa minha casa, minha vida**. 2016. 88 f. Dissertação (Mestrado em Administração) – Universidade de Brasília, Brasília, DF, 2016.